International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

MACHINE LEARNING ALGORITHMS IN BIG DATA ANALYTICS

K VIJAY KRUPA VATSAL[1], M PRAVEEN[2], M.ROOPA[3], DHANUSH KODI RANGA RAJAN[4], B NAVYA PRATHYUSHA[5], K VINEETH[6]

[1] Assistant Professor, Department of CSE, Malla Reddy Engineering College (Autonomous), Hyderabad, India

[2] Assistant Professor , Department of CSE, Malla Reddy Engineering College (Autonomous), Hyderabad, India

[3] Assistant Professor, Department of CSE, Malla Reddy Engineering College (Autonomous), Hyderabad, India

[4,5,6] UG Student, Department of CSE, Malla Reddy Engineering College (Autonomous), Hyderabad, India

Abstract- Big data is a wonderful supply of information and knowledge from the systems to other end-users. However handling such quantity of knowledge needs automation, and this leads to a trend of data processing and machine learning techniques. Within the ICT sector, as in several different sectors of analysis and trade, platforms and tools are being served and developed to assist professionals to treat their knowledge and learn from it automatically. Most of these platforms return from huge firms like Google or Microsoft, or from incubators at the Apache Foundation. This review explains Machine learning Algorithms in Big data Analytics, and machine learning challenges us to take decisions where there is no known "right path" for the specific problem based on previous lessons and enumerates some of the foremost used tools for analyzing and modeling big-data.

Keywords: Machine Learning Algorithms, Big data Analytics, Apache Foundation

I. Introduction

Machine learning (ML) systems have created enormous societal effects in an extensive variety of utilizations, for example, computer vision, speech processing, natural language understanding, neuroscience, healthcare, and Internet of Things. ML tends to the topic of how to assemble a process framework that enhances consequently through experience [1]. A ML issue is referred to as the issue of learning from past experience with respect to some tasks and performance measure. ML methods empower users to reveal hidden structure and make forecasts from extensive data sets. ML blossoms with proficient learning techniques, rich as well as vast data, and effective computing conditions. Figure 1 shows how machine learning is formed Big data has been described by five characteristics: volume (amount/measure of information), velocity (speed of information retrievals), variety (sort, nature, and arrangement of information), veracity (reliability/nature of caught information), and value (insights and effect). We

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

composed the five measurements into a stack, comprising of enormous information, and value layers beginning from the base. The lower layer (e.g., volume and speed) depends all the more intensely on mechanical advances, and higher layer (e.g., value) is more situated toward applications that load the key energy of enormous information. In order to understand the estimation of big data analytics and to process data such large effectively, existing ML standards and calculations should be adapted. As issues turn out to be progressively testing and requesting regularly tool boxes supporting ML programming improvement neglect to meet the desires regarding computational execution. Thus, the logical achievements without bounds will not including a doubt be controlled by cutting edge registering abilities that will permit analysts to control and investigate enormous datasets [2]. Therefore, ML has inconceivablen possibilities for, and is a fundamental piece of big data analytics [3]. Figure 1. Shows the process of generation of Machine learning. Machine Learning is a combination of Computer Science, Engineering and Statistics. Any field that needs to interpret and act on data can benefit from machine learning.

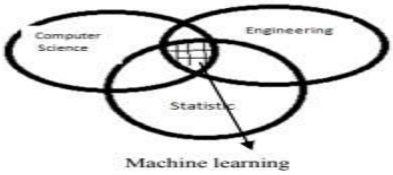

Figure 1 : Formation of Machine Learning

This paper mainly concentrates on ML procedures with regards to big data and modern computing environments. In particular, we mean to research openings and difficulties of ML on big data and big data exhibits new open doors for ML. For instance, big data empowers design learning at multi-granularity and decent variety, from different perspectives in a parallel manner. Also, big data gives chances to make causality deduction in view of chains of sequences. In many cases, big data also give major challenges to ML to extract perfect pattern from available test data. For example, data dimensionality, model versatility, distributed computing, stream of data, adaptability, and usability [20]. In this paper, we present a structure of ML on big data. The structure is focused on ML which takes after the periods of preprocessing, learning, and assessment. Section II contains survey of Big Data and Machine Learning. Section III contains different types of algorithms available in Machine Learning. Section IV describes role of machine learning algorithms in Big data. Section V discuss about Tools of machine learning in Big Data(Hadoop).

II. Big Data

Traditional data use centralized database architecture in which large and complex problems are solved by a single computer system. Centralized architecture is costly and ineffective to process large amount of data. Big data is based on the distributed database architecture where a large block of data is solved by dividing it into several smaller sizes. Then the solution to a problem

is computed by several different computers present in a given computer network. The computers communicate to each other in order to find the solution to a problem [2]. The distributed database provides better computing, lower price and also improve the performance as compared to the centralized database system. This is because centralized architecture is based on the mainframes which are not as economic as microprocessors in distributed database system. Also the distributed database has more computational power as compared to the centralized database system which is used to manage traditional data.

A. Types of data

Traditional database systems are based on the structured data i.e. traditional data is stored in fixed format or fields in a file. Examples of the structured data include Relational Database System (RDBMS) and the spreadsheets, which only answers to the questions about what happened. Traditional database only provides an insight to a problem at the small level. However in order to enhance the ability of an organization, to gain more insight into the data and also to know about metadata unstructured data is used [2]. Big data uses the semi-structured and unstructured data that improves the variety of the data gathered from different sources like customers, audience or subscribers. After the collection, Bid data transforms it into knowledge based information [3].

B. Storage & Cost

Under the traditional database system it is very expensive to store massive amount of data, so all the data cannot be stored. This would decrease the amount of data to be analyzed which will decrease the result's accuracy and confidence. While in big data as the amount required to store voluminous data is lower. Therefore the data is stored in big data systems and the points of correlation are identified which would provide high accurate results[11].

C. Machine Learning

Machine learning normally experiences information preprocessing, learning, and assessment stages Data preprocessing gets ready crude information into the "right form" for resulting learning steps. The raw data is probably going to be unstructured, loud, fragmented, and conflicting. The preprocessing step changes such information into a frame that can be utilized as contributions to learning through information cleaning, extraction, change, and combination. The learning stage picks learning calculations and tunes display parameters to produce wanted yields utilizing the preprocessed input information. Some learning techniques, especially authentic learning, can likewise be utilized for information preprocessing. The assessment takes after to decide the execution of the educated models. For example, execution assessment of a classifier includes dataset determination, execution measuring, mistake estimation, and factual tests [8]. The assessment results may prompt changing the parameters of picked learning calculations and additionally choosing diverse calculations. Designing a learning system, i.e. an application of

machine learning, involves four design choices. 1. Choosing the training data. 2. Choosing the target function. 3. Choosing the representation. 4. Choosing the learning algorithm.

III. Anatomy of Machine Learning

Managing enormous information normally includes finding there on the significant data to display the wear away or change the appearance or surface by long presentation to the information creating it, and redesigning it into accommodating data and information. For such objectives the greater part of unmistakable and prescient esteems are exceptionally helpful. Machine Learning strategies for demonstrating and expectation, information conglomeration and agglomeration, and information disclosure. Machine Learning, as a piece of information handling, gives techniques to treat and concentrate data from learning mechanically, wherever human administrators and experts aren't ready to touch upon because of the degree of multifaceted nature or the amount to be dealt with per measure. For quite a while, machine learning has been a science connected in hideously specific situations like medications, earth sciences or offering; however in view of the vision of enormous information everybody has the prerequisite to treat it and furthermore gain from it mechanically. As a reaction to the present, uncountable stages, devices, dialects and applications have appeared to treat that information where giving machine learning calculations some to steady preparing, some to locate phenomenal occasions in information, some for circulated conditions, however all spanning the yet existing hole amongst AI and examination and exchange. Such devices furthermore are fundamental perspectives for connected machine learning forms, aside of precision and multifaceted nature of calculations: viewpoints simply like the ability of parallelizing the preparation and expectation organizes, the kind of programming dialect to use for demonstrating our downside and recover our insight, available officially upheld libraries steady superior, or the approach the outcomes open territory measure intending to be gathered and shown Here we want to present a defense about the thoughts of machine learning and information preparing, and a few tools and stages used by information researchers to handle upon enormous information and perform machine learning forms
there on.

D. Machine Learning Algorithms with Data Analytics

Machine learning algorithms are described as learning a target function (f) that best maps input variables (X) to an output variable (Y).

$$Y = f(X)$$

This is a general learning task where we would like to make predictions in the future (Y) given new examples of input variables (X). We don't know what the function (f) looks like or its form. If we did, we would use it directly and we would not need to learn it from data using machine learning algorithms. The most common type of machine learning is to learn the mapping Y

= f(X) to make predictions of Y for new X. This is called predictive modeling or predictive analytics and the main aim is to make the most accurate predictions possible. If we use big data for storing bulk amount of information and manipulation, is one thing but extracting useful information from these will possible through machine learning. With this machine learning we can extract efficient patterns.

E. Parametric & Non-Parametric Machine Learning

Algorithms Assumptions can greatly simplify the learning process, but can also limit what can be learned. Algorithms that simplify the function to a known form are called parametric machine learning algorithms. The algorithms involve two steps: 1.Select a form for the function. 2. Learn the coefficients for the function from the training data. Some examples of parametric machine learning algorithms are Linear Regression and Logistic Regression.

Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data. Non-parametric methods are often more flexible, achieve better accuracy but require a lot more data and training time. Examples of nonparametric algorithms include Support Vector Machines, Neural Networks and Decision Trees. Bias are the simplifying assumptions made by a model to make the target function easier to learn. Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias. Decision trees are an example of a low bias algorithm, whereas linear regression is an example of a high-bias algorithm[12]. Variance is the amount that the estimate of the target function will change if different training data was used. The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance, not zero variance. The k- Nearest Neighbors algorithm is an example of a high- variance algorithm, whereas Linear Discriminate Analysis is an example of a low variance algorithm. The goal of any predictive modeling machine learning algorithm is to achieve low bias and low variance. In turn, the algorithm should achieve good prediction performance. The parameterization of machine learning algorithms is often a battle to balance out bias and variance. Increasing the bias will decrease the variance. Increasing the variance will decrease the bias. Predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explains ability. We will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

Linear regression is an equation that describes a line that best fits the relationship

between the input variables (x) and the output variables (y), by finding specific weightings for the input variables called coefficients (B).

$$y = B0 + B1 * x$$

We will predict y given the input x and the goal of the linear regression learning algorithm is to find the values for the coefficients B0 and B1.Different techniques can be used to learn the linear regression model from data, such as a linear algebra solution for ordinary least squares and gradient descent optimization. We have more than two classes then the Linear Discriminate Analysis algorithm is the preferred linear classification technique Some good rules of thumb when using this technique are to remove variables that are very similar (correlated) and to remove noise from the data, if possible. It is a fast and simple technique and good first algorithm. The representation of LDA consists of statistical properties of the data, calculated for each class. For a single input variable this includes: The mean value for each class. The variance calculated across all classes. Predictions are made by calculating a discriminate value for each class and making a prediction for the class with the largest value. The technique assumes that the data has a Gaussian distribution (bell curve), so it is a good idea to remove outliers from data. It's a simple and powerful method for classification\predictive modeling problems. Decision tree model is a binary tree representation model. Each node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. Predictions are made by walking the splits of the tree until arriving at a leaf node and output the class value at that leaf node. Trees are fast to learn and very fast for making predictions [18]. They are also often accurate for a broad range of problems and do not require any special preparation for the data. Decision trees have a high variance and can yield more accurate predictions when used in an ensemble. Naive Bayes is a simple but surprisingly powerful algorithm for predictive modeling. The model is comprised of two types of probabilities that can be calculated directly from training data: The probability of each class. The conditional probability for each class is given for each x value. Once calculated, the probability model can be used to make predictions for new data using Bayes Theorem[16]. When the data is real-valued it is common to assume a Gaussian distribution (bell curve) so that we can easily estimate these probabilities. Naive Bayes is called naive because it assumes that each input variable is independent. This is a strong assumption and unrealistic for real data, nevertheless, the technique is very effective on a large range of complex problems. K-Nearest Neighbor predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression problems, this might be the mean output variable, for

classification problems this might be the mode (or most common) class value. The trick is in how to determine the similarity between the data instances. The simplest technique if the attributes are all of the same scale (all in inches for example) is to use the Euclidean distance, a number we can calculate directly based on the differences between each input variable.KNN can require a lot of memory or space to store all of the data, but only performs a calculation (or learn) when a prediction is needed, just in time. we can also update and accurate the training instances over time to keep predictions accurate. The idea of distance or closeness can break down in very high dimensions (lots of input variables) which can negatively affect the performance of the algorithm on the problem. This is called the curse of dimensionality [17]. It suggests only those input variables that are most relevant to predicting the output variable. Linear Vector Quantization is an artificial neural network algorithm that allows choosing how many training instances to hang onto and learns exactly the final instances. By grouping input sequences together and encoding them as a single block, we can obtain efficient loss as well as lossless compression algorithms. There are also some quantization techniques that operate on blocks of data. We can look at these blocks as vectors. This kind of quantization technique is called vector quantization [4]. The representation for LVQ is a collection of codebook vectors. These are selected randomly in the beginning and adapted to best summarize the training dataset over a number of iterations of the learning algorithm. After learned, the codebook vectors can be used to make predictions just like K-Nearest Neighbors. The most similar neighbor (best matching codebook vector) is found by calculating the distance between each codebook vector and the new data instance. The class value (real value in the case of regression) for the best matching unit is then returned as the prediction. Best results are achieved if you rescale you're to have the same range, such as between 0 and 1[15].

Support Vector Machines learning algorithm finds the coefficients those results in the best separation of the classes by the hyper plane. The distance between the hyper plane and the closest data points is referred to as the margin. The best or optimal hyper plane that can separate the two classes is the line that has the largest margin. Only these points are relevant in defining the hyper plane and in the construction of the classifier. These points are called the support vectors. They support or define the hyper plane and optimization algorithm is used to find the values for the coefficients that maximize the margin. SVM might be one of the most powerful out-of-the-box classifiers and worth trying on the dataset [14]. Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks .A property of deep learning is that the performance of this type of model improves by training them with more examples by increasing their depth or representational capacity. In

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

addition to scalability, another often cited benefit of deep learning models is their ability to perform automatic feature extraction from raw data, also called feature learning[21] Transfer learning is an assumption of traditional machine learning methodologies is the training data and testing data are taken from the same domain, such that the input feature space and data distribution characteristics are the same. However, in some real-world machine learning scenarios, this assumption does not hold. There are cases where training data is expensive or difficult to collect. Therefore, there is a need to create high-performance learners trained with more easily obtained data from different domains. This methodology is referred to as transfer learning. [22]defines transfer learning, presents information on current solutions, and reviews applications applied to transfer learning. The transfer learning solutions surveyed are independent of data size and can be applied to big data environments.

Figure 2: Algorithms in machine learning

IV. Machine Learning Algorithms in Big Data Analytics

Machine Learning is a sub-field of data science that focuses on designing algorithms that can learn from experience [5]and make predictions on the data. A computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance as tasks T, as measured by P improves the experience E. Machine learning experience includes supervised learning, Unsupervised Learning and Reinforcement Learning methods[7]. Unsupervised methods actually start off from unlabeled data sets, so, in a way, they are directly related to finding out unknown properties in them (e.g. clusters or rules). Machine learning focuses on prediction [8], based on known properties learned from the training data. Data mining (which is the analysis step of Knowledge Discovery in Databases) focuses on the discovery of (previously) unknown properties on the data .For instance, performance evaluation of a classifier involves dataset selection, performance measuring, error-estimation, and statistical tests [6]. The evaluation results may lead to adjusting the parameters of chosen learning algorithms and/or selecting different algorithms. While productive uses of machine learning can't depend entirely on packing consistently expanding measures of massive Information at calculations and seeking after the best, the capacity to use a lot of information for machine learning tasks is a categorical requirement has ability for specialist now. While much of machine

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

learning holds true regardless of data amounts, there are aspects which are the exclusive domain of Big Data modeling, or which apply more so than they do to smaller data amounts. Figure 1 outlines a process for applying machine to Big Data in his original graphic. The process includes paths for descriptive, predictive, and prescriptive analysis, as well as
simulation. Importantly, the machine learning process is explicitly noted as recursive, which is perhaps especially true of modeling large quantities of data, and it also breaks down the relative number of records at each successive stage of a machine learning task.

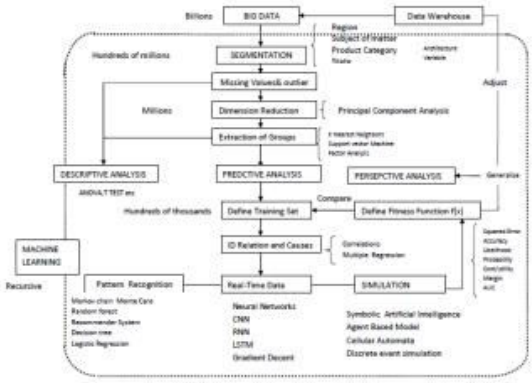

Figure 3: Role of Machine Learning in Big Data

F. Cohort of Data sets in Machine Learning

The performance measure is very much necessary to evaluate an algorithm in machine learning. The performance measure on unseen data set is known as test set which is different from data set. Data set which trains itself is called as training set. Data sets are in a perfect world produced from a probability distribution. While creating the data we typically accept that the awareness in the data set is autonomous from each other additionally that they are indistinguishably circulated. This presumption is otherwise called autonomous and indistinguishably distributed (a.i.d).The training and test data set for the calculations fundamentally need to originate from the same probability distribution and are a.i.d. Having parceled our data set into training set and test sets, we prepare the learning calculations by utilizing the training set in order to diminish the training error. We at
that point apply the "trained" model to the test set which is inconspicuous information. The test error is constantly more prominent than the training error. In this manner we can decrease the training error and in addition diminish the contrast between training error and test error.

V. Tools for Machine Learning Algorithm in Big Data Analytics

In Big-Data situations, operators, managers and information researchers need to acquire data and learning from immense data sets or from wide surges of data. Keeping in mind the end goal to make this procedure simple, letting information researchers to concentrate on mechanizing this procedure and concentrate on the outcomes, a few systems have seemed to give such administration. Here a couple of them, however not by any means the only ones, are abridged.

G. Map-Reduce frameworks:

Apache Hadoop and Spark Most machine learning procedures can be parallelized by understanding a calculation procedure for each information and after that total the procedure yields into the arrangement. Such

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

procedures can be explained utilizing a Map- Reduce arrangement: data is split in parts, each part is processed in parallel and the outcomes are amassed into the arrangement. Apache Hadoop [26] is a generally utilized open-source structure in Java for such purposes. Clients can send without anyone else bunches or server farms, or can get an answer from organizations offering Hadoop as Platform as a Service and concentrating just on conveying their

applications. In view of Hadoop, Apache Spark [27] is an answer endeavoring to enhance execution by centering in particular functionalities like machine learning, graph analysis and data-streaming, and programmable in Scala or Python. While Hadoop has been available for long time and as of now has more capacities covering more parts of the business, Spark and Mahout develops by centering and enhancing particular issues, the vast majority of them related with machine learning and enormous information treatment.

H. Apache Spark

Apache Spark is a general-purpose analytics framework. It improves efficiency through in-memory computing primitives, Pipelined computation and it improves usability through APIs in Scala, but Java, Python, and R APIs also available and also works through interactive Shell. Spark provides a general middleware layer that re-implements existing learning tasks so they can run on a big data platform. Such a middleware layer often provides general primitives/operations that are useful for many learning tasks. This approach is suitable for users to try different learning tasks/algorithms within the same framework. The other category is to transform individual learning algorithms to run on a big data platform.

I. Apache Mahout:

Mahout is an open source project from Apache, offering Java libraries for distributed or otherwise scalable machine-learning algorithms. Popular implementation of Mahout will be done in the fashion of Latent Dirichlet Allocation. Mahout implementation of LDA are Collapsed Variation Bayes (CVB) and pipeline of Hadoop jobs.

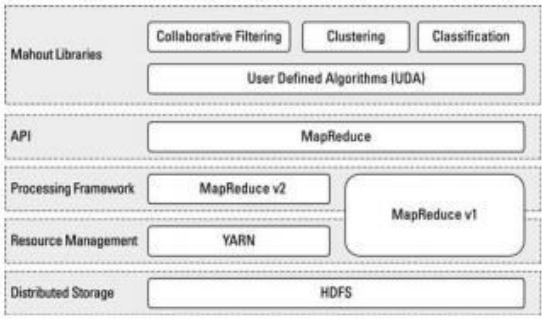

Figure 4: Framework of Apache Mahout

These calculations cover exemplary machine learning undertakings, for example, arrangement, bunching, affiliation run examination, and suggestions. In spite of the fact that Mahout libraries are intended to work inside an Apache Hadoop setting, they are likewise good with any framework supporting the Map Reduce system. For instance, Mahout gives Java libraries to Java accumulations and basic math operations (direct variable based math and measurements) that can be utilized without Hadoop[23]. Many machine learning

algorithms are supported by Spark MLlib and Apache Mahout, because they are separated as front-end and back-end layers for algorithms and execution engines. It has a compatibility to move from big data engine to other.

VI. Conclusion

. ML is fundamental to address the difficulties postured by big data and reveal concealed patters, information, and bits of knowledge from enormous information keeping in mind the end goal to transform the capability of the last into genuine incentive for business basic leadership and logical investigation. Future scope of Machine learning analytics is

how to make ML more declarative, so that it is easier for non- experts to specify and interact with different type of data in different streams. In the future, we will enhance and assess the performance of machine learning techniques for different types of problems. One promising direction is to extend the machine learning approaches towards big data, which are efficient and highly scalable in the way they process high-dimensional data. From this survey perspective, we intend to maintain focus on the land utilization and land classification (LULC) application and explore the benefits of using different classifications, clustering and prediction. We will also study the effect of different spatial resolutions on the performance of these machine learning techniques and moreover given techniques in enhancing its entry for the diverse e-science zones with various sectors.

.References

[1]. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, pp. 255-260, 2015.

[2]. Hey, T., Tansley, S., Tolle, K., editors (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. S.Hasan et a.l

[3]. Sun, Y. et al., 2014. Organizing and Querying the Big Sensing Data with Event-Linked Network in the Internet of Things. International Journal of Distributed Sensor Networks, 14, p.11.

[4]. Fan, J., Han, F. & Liu, H., 2014. Challenges of Big Data analysis. National Science Review , 1 (2), pp.293– 314.

[5]. Parmar, V. & Gupta, I., 2015. Big data analytics vs Data Mining analytics. IJITE, 3(3), pp.258–263.

[6]. K. Sayood, Introduction to Data Compression, Morgan Kaufinarm Publishers, San Francisco, CA, 2000.

[7]. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, pp. 255-260, 2015.

[8]. N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective: Cambridge University Press, 2011.

[9]. S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed.: Prentice Hall, 2010.

10]. Y.Bengio ,A.Courville, and P. Vincent, "Representation learning: A review and new perspectives," Pattern Analysis and Machine

Intelligence, IEEE Transactions on, vol. 35, 2013.